

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 05-265860

(43)Date of publication of application : 15.10.1993

(51)Int. CI. G06F 12/08

G06F 3/06

(21)Application number : 04-064187

(71)Applicant : HITACHI LTD
HITACHI SOFTWARE ENG CO
LTD

(22)Date of filing : 19.03.1992

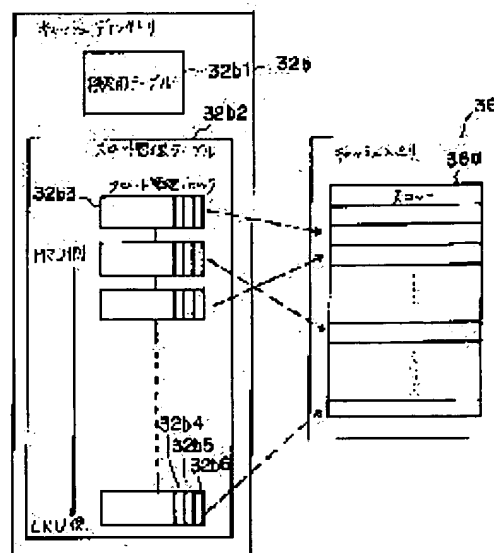
(72)Inventor : TANAKA MAYUMI
ISAKA SHOJI
MORI EIKI
YOTSUYA MORIHIKO
HONMA SHIGEO

(54) DISK CACHE INPUT/OUTPUT CONTROL SYSTEM

(57)Abstract:

PURPOSE: To transfer data between a cache memory and a host processor without waiting for the destage end of a slot in the destage when an input/output request is generated from the host processor to the slot.

CONSTITUTION: This disk cache input/output control system is provided to request access from a high-order channel parallel to the destage even when a slot 36a in a cache memory 36 containing data, to which the access is requested from the high-order channel, is under destaging by setting slot managing information 1 storage part-3 storage part 32b4-32b6 concerning the destage to a slot managing table 32b2 in a cache directory 32b. Therefore, since data can be transferred between the host processor and the cache memory without waiting for the destage end, the throughput of the system can be prevented from being lowered.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998, 2003 Japan Patent Office

*** NOTICES ***

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

CLAIMS

[Claim(s)]

[Claim 1] It has cache memory with two or more slots holding the counterpart of the data stored in the rotation mold store. In the disk cache equipment which answers using the data currently held at said cache memory on the occasion of activation of the input/output instruction to said rotation mold store given from a host processor The 1st transfer control means which performs data transfer between said cache memory and said host processors, The 2nd transfer control means which performs data transfer between said cache memory and said rotation mold store, A storage means to correspond to each slot of said cache memory, and to store the management information about each corresponding slot, Write said management information in said storage means, and said management information stored in said storage means at the time of activation of said input/output instruction is referred to. It judges whether it is data with which the data stored in the corresponding slot do not bar the writing to said rotation mold store by said 2nd transfer control means. It is the disk cache input/output control system characterized by having the processor which publishes the command for transfer control to said 1st transfer control means about the data which do not bar writing.

[Claim 2] In the disk cache input/output control system of the claim 1 above-mentioned publication, said management information corresponding to each slot of said cache memory The data stored in each corresponding slot are written in said rotation mold store (DESUTEJI processing). Whether it is under ***** The data stored in which record in a correspondence slot and in the inside of DESUTEJI And disk cache input/output control system characterized by being the information on whether writing was performed by activation of said input/output instruction into DESUTEJI in the data stored in the correspondence slot at said data.

[Claim 3] When the data to the write instruction given from said host processor exist in said cache memory in the disk cache input/output control system of the claim 2 above-mentioned publication Said management information stored in said storage means by said processor is read. Even if it judges whether it is among DESUTEJI from said cache memory to said rotation mold store about said data by this management information and said 2nd transfer control means is DESUTEJI [from said cache memory to said rotation mold store] processing About the data with which DESUTEJI processing was ended, it writes in to said data in said cache memory immediately. Disk cache input/output control system characterized by reporting write-in termination to said host processor after storing in said management information the information that it wrote in to said data in said cache memory during DESUTEJI processing.

[Translation done.]

*** NOTICES ***

JPO and NCIPi are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3. In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Industrial Application] This invention is applied to the disk cache input/output control system in the disk subsystem in an information processor, and relates to an effective technique.

[0002]

[Description of the Prior Art] In information processing system, a central processing unit has the high-speed primary storage which consists of semiconductor memory, and the processing speed is improving remarkably in recent years. Using the magnetic disk drive which is a kind of a rotation mold store as external storage on the other hand is known.

[0003] By the way, the magnetic disk drive has structure which piled up the record disk (this is called a magnetic disk) of two or more sheets. As a field which records data, on a concentric circle, two or more trucks are formed and two or more records are stored on each magnetic disk at one truck. The truck which is in the equal distance from on the revolving shaft of a magnetic disk is perpendicularly located in a line by only the number of sheets of a disk. This set is called the cylinder.

[0004] If a cylinder number, a track number, and a record number are specified when accessing data The magnetic head attached on the field of each magnetic disk a predetermined driving gear Make it move in the direction of a path of a magnetic disk, and the record of positioning and the purpose is recorded on the truck of arbitration. Or when the field on the truck concerned where a record should be recorded passes directly under the magnetic head by rotation of a magnetic disk, the writing of read-out or a record is performed for the target record.

[0005] Therefore, mechanical movements, such as rotational delay until the record of the purpose on the seek operation which moves the magnetic head to the target truck, or the truck concerned passes directly under the magnetic head, will intervene. Since it is large as compared with the processing time of a central processing unit, the time amount which such mechanical movement takes is one factor in which the I/O time of the data from a magnetic disk presses down the engine performance of information processing system.

[0006] The so-called disk cache technique it was made to answer using the data which arrange the cache memory which consists of semiconductor memory, copy beforehand the high data of possibility that I/O will be required from a host-processor side to cache memory from the magnetic disk drive, and are copied as much as possible for the data transfer path between a host processor and a magnetic disk drive by the cache memory concerned to the input/output request from a host-processor side that it should avoid above un-arranging is used. Thereby, the target data are transmitted between cache memory and a host processor, the mechanical movement to a magnetic disk drive of them is lost, and the high-speed response of them is attained.

[0007] Generally, with a disk cache technique, if the target record exists in cache memory to the input instruction from a host processor, data transfer will be performed from cache memory to a host

processor, and improvement in the speed of a response will be attained.

[0008] On the other hand, to the output instruction from a host processor, there are two kinds of arts called a write-through mold disk cache and a light after mold disk cache.

[0009] In a write-through mold disk cache, since cache memory is volatility, the output data transmitted from a host processor write in both cache memory and a magnetic disk drive, and prevent the data loss at the time of power-source cutoff etc. For this reason, the mechanical movement to a magnetic disk drive intervenes, and improvement in the speed of a response cannot be attained.

[0010] On the other hand, by the light after mold disk cache, the un-volatilizing-ized device other than cache memory is provided, and when the write-in data transmitted from a host processor are written in both cache memory and an un-volatilizing-ized device, they report write-in termination to a host processor, for example, as indicated by JP,59-220856,A etc. Since write-in data are backed up in the un-volatilizing-ized device, the data guarantees at the time of power-source cutoff etc. can be offered, the mechanical movement to a magnetic disk drive does not intervene at the time of writing, but they can attain improvement in the speed of a response.

[0011] Now, with a disk cache technique, although the probability for the data of a hit and the purpose to exist in cache memory that the target data exist in cache memory is called a hit ratio, in order to make [many / as possible] the rate of the data transfer between cache memory and a host processor, it is required to enlarge a hit ratio.

[0012] The LRU (Least Recently Used) algorithm is widely known as an algorithm which chooses the data which should be stored in cache memory. It thinks that possibility of a LRU algorithm that the data accessed most recently will continue to be accessed is high, is more long, and is the algorithm which it is going to stop in cache memory.

[0013] Therefore, when the data demanded from the host processor do not exist in cache memory (this is called a mistake), some whole trucks or the whole truck containing the data is copied to cache memory from a magnetic disk drive (this is called a staging). When it is going to carry out a staging and there is no free area into cache memory, by the LRU algorithm, the data accessed in the oldest past think that possibility of being accessed is the lowest, cancel out of cache memory, and newly perform a staging to this field from now on.

[0014] By the way, in the light after mold disk cache which stores the counterpart of data also in an un-volatilizing-ized device at cache memory and coincidence in order to prevent loss of write-in data at the time of failures, such as emergency interruption of service, since an un-volatilizing-ized device is small capacity, data are copied to a magnetic disk drive at any time from cache memory so that it may not overflow with write-in data (this is called DESUTEJI). The data [DESUTEJI / data] are canceled from an un-volatilizing-ized device. DESUTEJI [disk storage control / disk storage control uses idle time and / the whole truck containing the lowest data of possibility of being accessed by the above-mentioned LRU algorithm from now on, or two or more trucks near the truck concerned] at any time. Or DESUTEJI [above-mentioned DESUTEJI / it does not catch up, and / the truck containing the write-in data concerned, and two or more trucks in the near] when the write-in data more than a certain constant rate are already stored and it may be full of an un-volatilizing-ized device when there is a demand of the writing from a host processor and it is going to write write-in data in cache memory and an un-volatilizing-ized device.

[0015] Thus, DESUTEJI [with disk storage control / it is managed efficiently and / the written-in data / cache memory / a magnetic disk drive] at any time so that the un-volatilizing-ized device of small capacity may not overflow with write-in data. Moreover, it is also possible to perform DESUTEJI positively with the direct instruction from a host processor.

[0016]

[Problem(s) to be Solved by the Invention] Although cache memory consists of two or more entries (this entry is called a slot) and the data for one truck of a magnetic disk drive are stored in each entry with the above conventional techniques, when input/output instruction is published from a host processor and the

truck (slot) containing the data concerned copied in cache memory is among DESUTEJI, the slot concerned becomes under use (it is called slot BIJI). For this reason, the input/output instruction from a host processor will be kept waiting until DESUTEJI of the slot concerned is completed. In spite of the target data existing in cache memory (hit), the data transfer between cache memory and a host processor will be kept waiting by DESUTEJI, and there was a problem that the availability of a system throughput (the amount of data transfer per unit time amount) and a data transfer path fell.

[0017] Then, the purpose of this invention is to offer the input/output control technique which makes possible data transfer between cache memory and a host processor, even if the slot of the cache memory containing the data which had input/output request from the host processor is among DESUTEJI, in order to prevent the fall of a throughput, and the fall of the availability of a data transfer path.

[0018] The other purposes and the new description will become clear from description and the accompanying drawing of this specification along [said] this invention.

[0019]

[Means for Solving the Problem] It will be as follows if the outline of a typical thing is briefly explained among invention indicated in this application.

[0020] namely, in the disk cache input/output control system which becomes this invention The cache memory which holds the copy of data delivered and received among both between a host processor and a magnetic disk drive, especially a light after mold disk cache (by the data write request published from a host processor) The cache memory which performs to asynchronous actuation which stores in cache memory the write-in data which should be written in a magnetic disk drive, and actuation written in a magnetic disk drive from the cache memory concerned is prepared. Activation of the input/output instruction published from a host processor is faced. It is the information processing system it was made to answer as much as possible using the data currently held at cache memory. The slot management information concerning DESUTEJI to each slot in cache memory, 1. -- the slot concerned -- the inside of DESUTEJI ***** -- 2. -- the slot concerned at the time in DESUTEJI which record of the slot concerned -- the inside of DESUTEJI, or 3. -- in DESUTEJI the slot concerned When a means to memorize whether it wrote in the slot concerned is established and input/output request is published from a host processor, even if the slot containing the target data is among DESUTEJI The input/output request concerned is received and it makes it possible to perform a host processor, the data transfer between cache memory, and DESUTEJI to juxtaposition.

[0021]

[Function] as the slot management information 1 of the above-mentioned when DESUTEJI occurs, for example to the slot in cache memory according to the disk cache input/output control system of above-mentioned this invention -- " -- the information [slot / concerned] "in DESUTEJI is memorized. And in case DESUTEJI [the record currently held in the slot concerned] one by one, the number of the record which starts DESUTEJI is memorized as the above-mentioned slot management information 2. When input/output request is published from a host processor, the demand concerned judges a read-out demand or a write request. It is a read-out demand, and when the target data exist in cache memory, the above-mentioned slot management information 1 investigates whether it is among DESUTEJI about the slot containing the target data, and even if it is among DESUTEJI, the target data are immediately transmitted to a host processor from cache memory.

[0022] On the other hand, the input/output request from a host processor is a write request, and it investigates whether it is among DESUTEJI about the slot which contains the target data by the slot management information 1 like [when the target data exist in cache memory] the case of an above-mentioned read-out demand, and in being among DESUTEJI, the above-mentioned slot management information 2 investigates the record number within the slot in current DESUTEJI. If DESUTEJI of the record containing the target data is completed, immediately, the data from a host processor will be copied to the record of the purpose within the slot concerned, the record concerned will be copied also to a writing and un-volatilizing-ized device, and write-in termination will be reported to a

host processor. To coincidence, what "was written in the slot concerned into DESUTEJI in the slot concerned" is memorized as the above-mentioned slot management information 3. This prevents that the slot concerned is canceled from an un-volatilizing-ized device after DESUTEJI completion of the slot concerned.

[0023] In addition, the information memorized by the three above-mentioned slot management information shall be reset after DESUTEJI completion.

[0024] When input/output request is published from a host processor by doing in this way, even if the slot in the cache memory containing the target data is among DESUTEJI In the time of the write request to the read-out demand and the record which DESUTEJI has already ended from high order equipment In parallel to cache memory and DESUTEJI performed using the low order pass which is a data transfer path between magnetic disk drives, data transfer using the high order pass which is a data transfer path between cache memory and a host processor can be performed. For this reason, a data transfer path is used effectively, without the data transfer between a host processor and cache memory waiting for DESUTEJI termination, it performs immediately and a system throughput improves.

[0025]

[Example] Hereafter, an example of the disk cache input/output control system which is one example of this invention is explained to a detail, referring to a drawing.

[0026] Drawing 1 is the block diagram showing an example of the whole configuration of the information processing system with which disk cache input/output control of this invention is carried out.

[0027] As shown in drawing 1 , the information processing system with which disk cache input/output control of this example is carried out consists of central processing units 1a and 1b, channel 2a, 2b, light after mold disk cache equipment 3, the magnetic-disk contact 4 and two or more magnetic disk drives 5, the high order pass 6, and the low order pass 7.

[0028] Channel 2a and 2b control the command between central processing units 1a and 1b and light after mold disk cache equipment 3, and transfer of data.

[0029] By the command of light after mold disk cache equipment 3, the magnetic-disk contact 4 chooses suitably two or more magnetic disk drives 5 of a subordinate, and operates connecting with the light after mold disk controller concerned etc.

[0030] A magnetic disk drive 5 is external storage which records / reproduces said channel 2a and the data which are delivered and received between 2bs through light after mold disk cache equipment 3 and the magnetic-disk contact 4.

[0031] The high order pass 6 has connected channel 2a, 2b, and light after mold disk cache equipment 3, and data transfer between channel 2a, 2b, and light after mold disk cache equipment 3 is performed through this pass.

[0032] the low order pass 7 -- light after mold disk cache equipment 3 and the magnetic-disk contact 4 -- connecting -- **** -- this pass -- minding -- ** -- data transfer between light after mold disk cache equipment 3 and the magnetic-disk contact 4 is performed.

[0033] Drawing 2 is the block diagram showing an example of the detailed configuration of the light after mold disk cache equipment 3 of drawing 1 .

[0034] Light after mold disk cache equipment 3 consists of a microprocessor 31, a control memory 32, the data transfer control circuit 33, the transfer control circuit 34 for a channel, the transfer control circuit 35 for a disk, cache memory 36, and the un-volatilizing-ized device 37, as shown in drawing 2 .

[0035] A microprocessor 31 controls the light after mold disk cache control unit 3 whole. The control memory 32 stores micro program 32a and cache directory 32b.

[0036] The data transfer control circuit 33 controls the data transfer between channel 2a, 2b, and a magnetic disk drive 5, taking the synchronization with the transfer control circuit 34 for a channel, and the transfer control circuit 35 for a disk.

[0037] The transfer control circuit 34 for a channel controls transfer of the command between channel 2a

and 2b, data, etc., and the transfer control circuit 35 for a disk controls transfer of the command between magnetic disk drives 5, data, etc.

[0038] The cache memory section 36 consists of volatile semiconductor memory, and usually has two or more slot 36a. The counterpart of the data for one truck currently recorded on the magnetic disk drive 5 is stored in each slot 36a.

[0039] On the other hand, the un-volatilizing-ized device 37 is for consisting of semiconductor memory of a non-volatile etc., holding write-in data certainly also at the time of failures, such as emergency interruption of service, and securing the dependability of the write-in data concerned in order to guarantee writing data in a magnetic disk drive 5 certainly.

[0040] Directory memory 32b in the control memory 32 of drawing 2 consists of the slot managed table 32b1 and 32b2 for retrieval, as shown in drawing 3.

[0041] The table 32b1 for retrieval holds the address of the truck in a magnetic disk 5 of the data stored in cache memory 36 (staging), and the information which shows the storing location in the cache memory 36 of the data. A microprocessor 31 judges whether the data demanded from central processing unit 1a or 1b exist in cache memory 36 by referring to this table 32b1 for retrieval.

[0042] The slot managed table 32b2 holds each slot 36a of cache memory 36, and the entry corresponding to 1 to 1. Each entry is for being set in order according to the above-mentioned LRU algorithm, consisting of semiconductor memory of a non-volatile etc., writing in the side in which the data accessed most recently are located also at the time of failures, such as emergency interruption of service, in order to guarantee AKUSE ***** the MRU (Most Recently Used) side in a call and the oldest past, holding data certainly, and securing the dependability of the write-in data concerned.

[0043] Directory memory 32b in a control memory 32 consists of a slot managed table 32b1 and 32b2 for retrieval, as shown in drawing 3. The table 32b1 for retrieval holds the information which shows the storing location in the address and the cache memory 36 of data of the truck in the magnetic disk drive 5 of the data stored in cache memory 36 (staging). A microprocessor 31 judges whether the data demanded from central processing unit 1a or 1b exist in cache memory 36 by referring to this table 32b1 for retrieval.

[0044] The slot managed table 32b2 holds each slot of cache memory 36, and the entry corresponding to 1 to 1. Each entry is set in order according to the above-mentioned LRU algorithm, and calls the side in which the data accessed the MRU (Most Recently Used) side in a call and the oldest past are located in the side in which the data accessed most recently are located the LRU (Least Recently Used) side.

[0045] When canceling the data (data for one truck currently recorded on the magnetic disk drive 5) stored in one slot 36a in cache memory 36, the truck corresponding to the entry in the LRU side is chosen.

[0046] The information memorized to one entry is the magnetic disk drive number of the data stored in corresponding slot 36a, for example, a cylinder number, a track number, etc. Moreover, the three above-mentioned slot management information 1-3 by this invention is stored in the slot management block 32b3 at the slot management information 1 storing section 32b4 - the slot management information 3 storing section 32b6, respectively.

[0047] Read-out / write-in actuation of the data between cache memory 36 and the un-volatilizing-ized device 37, and a magnetic disk drive 5 are performed considering the amount of data for one truck in a magnetic disk drive 5 etc. as a unit, and transfer of the data between channel 2a of a high order or 2b is usually performed to asynchronous.

[0048] Next, an operation of the disk cache input/output control system of this example is explained, referring to the flow chart of drawing 4 and drawing 5. Drawing 4 is the command of MPU31 of the light after mold disk kish equipment 3 of drawing 2, and is a flow chart which shows the data transfer processing performed with the high order pass 6 by the transfer control circuit 34 for a channel. The data area where access by channel 2a and 2b is expected among the data stored in the magnetic disk drive 5 is copied to cache memory 36 by MPU31 at any time per truck etc. To channel 2a or the read-out demand

from 2b, a high speed is answered as much as possible using the data copied to the cache memory 36 concerned. To channel 2a or the write request from 2b, it writes in the data copied to cache memory 36, and writes also in the un-volatilizing-ized device 37 at coincidence. Then, write-in termination is immediately reported to channel 2a of data write request issue origin, or 2b. DESUTEJI [two or more data written in the un-volatilizing-ized device 37 / at any time / the target magnetic disk drive 5] collectively. It is canceled from the un-volatilizing-ized device 37, the data area concerned within the un-volatilizing-ized device 37 is opened wide, and the data written in the magnetic disk drive 5 are used for other write-in data.

[0049] By the way, since the slot concerned is using it when the magnetic disk drive 5 is DESUTEJI [device / 37 / un-volatilizing-ized] and an access request occurs from high order channel 2a or 2b to the data area in current and DESUTEJI (slot) conventionally, high order channel 2a or the access request to the slot concerned from 2b is not received until DESUTEJI is completed. In such a case, it controls by this example as follows using the slot management information 1, 2, and 3 respectively stored in the slot management information 1 storing section 32b4 within the slot management block 32b3 in the slot managed table 32b2, the slot management information 2 storing section 32b5, and the slot management information 3 storing section 32b6 to be able to process DESUTEJI processing, high order channel 2a, or the access request from 2b to juxtaposition.

[0050] That is, MPU31 of the light after mold disk cache equipment 3 of drawing 2 investigates whether the counterpart of the data by which the access request was carried out from high order channel 2a or 2b is stored in cache memory 36 by investigating the table 32b1 (referring to drawing 3) for retrieval in a cache directory 32 (step 101). When the counterpart of the target data is not stored in cache memory 36, processing (processing of a cache mistake) by the conventional technique is performed (step 110). When the counterpart of the target data is stored in cache memory 36, refer to the slot management block 32b3 which has managed the slot containing the counterpart of the data concerned from the slot managed table 32b2 in cache memory 36 for MPU31. MPU31 investigates first whether the slot concerned is among DESUTEJI using the information stored in the slot management information 1 storing section 32b4 (step 102). When the slot concerned is not among DESUTEJI, processing (cache hit) by the conventional technique is processed (step 107,108,109). When the slot concerned is among DESUTEJI, MPU31 judges whether high order channel 2a or the access request from 2b is read-out, or it is writing (step 103). When high order channel 2a or the access request from 2b is a read-out demand, by the command of MPU31, using the high order pass 6, immediately, the data transfer control circuit 34 for a channel reads from the slot in DESUTEJI in cache memory 36, and transmits data to high order channel 2a of demand issue origin, or 2b (step 108). Thereby, processing of a lead hit is performed, without waiting for termination of DESUTEJI to high order channel 2a to the slot in DESUTEJI, or the read-out demand from 2b.

[0051] When high order channel 2a or the access request from 2b is a write request MPU31 was stored in the slot management information 2 storing section 32b5 within the slot management block 32b3 which manages slot 36a containing the counterpart of the target data. It compares with the record number (referred to as record-number b) in the slot 36a concerned containing the data which had 2b to drawing and high order channel 2a or an access request in the record number in DESUTEJI in the slot 36a concerned (referred to as record-number a) (step 104). Generally, since DESUTEJI is performed in an order from the head record within a slot, if it is record-number a > record-number b, DESUTEJI will already have ended the data with an access request from high order channel 2a or 2b. At this time, it writes in the data of record-number b of the slot concerned in cache memory 36 immediately using the high order pass 6 by the data transfer control circuit 34 for a channel by the command of MPU31, and light hit processing which writes the same data also in the un-volatilizing-ized device 37 at coincidence is performed (step 105), and write-in termination is reported to high order channel 2a of data write request issue origin, or 2b. Furthermore, since MPU31 wrote in to record-number b in the slot 36a concerned into DESUTEJI, In order to prevent that the counterpart of the data within the slot concerned

is canceled from the un-volatilizing-ized device 37 after DESUTEJI termination of the slot 36a concerned, It records that the slot concerned had new writing in the slot management information 3 storing section 32b6 within the slot management block 32b3 in DESUTEJI (step 106).

[0052] Thus, high order channel 2a to the slot in DESUTEJI or the write request from 2b can be immediately processed as a light hit, without already waiting for termination of DESUTEJI of the whole slot, if it is an access request to the part which DESUTEJI ended. however, the data which had the access request from high order channel 2a or 2b when it was record-number $a \leq$ record-number b -- the inside of current and DESUTEJI -- or since being contained in the record [DESUTEJI / record] from now on is shown, it writes in by waiting for DESUTEJI termination of the slot concerned like the conventional technique (step 111).

[0053] In addition, being given as parameters, such as the Locate command in the record of what position within a slot whose the data by which the access request was carried out from whether high order channel 2a or the access request from 2b is read-out or it is writing, high order channel 2a, and 2b are contained it is a channel command, is known widely.

[0054] On the other hand, DESUTEJI processing with the low order pass 7 by the transfer control circuit 35 for a disk is explained using the flow chart of drawing 5 . MPU31 of drawing 3 supervises termination of DESUTEJI of the slot concerned, the record number which performs DESUTEJI at any time now about the slot of the cache memory 36 to which DESUTEJI is performed. First, it investigates whether DESUTEJI of the data for one slot was completed (step 201). When DESUTEJI of the data for one slot is not completed, it investigates whether DESUTEJI of one record of the arbitration within a slot was completed (step 202). When DESUTEJI for one record is not completed, DESUTEJI is continued as it is (step 204). When DESUTEJI for one record is completed, it records on the slot management information 2 storing section 32b5 of the slot management block 32b3 which has managed the slot concerned registered into the slot managed table 32b2 in cache memory 36 in the record number within the slot concerned which starts DESUTEJI next (step 203), and the data transfer control circuit 35 for a disk performs DESUTEJI of the record concerned (step 204).

[0055] When DESUTEJI for one slot is completed, MPU31 investigates whether the slot concerned was written in to the slot concerned into DESUTEJI by high order channel 2a or the write request from 2b with reference to the slot management information 3 storing section 32b6 of the slot management block 32b3 which has managed the slot concerned (step 205). When it does not write in, the counterpart of the data of the slot concerned is canceled from the un-volatilizing-ized device 37 (step 206). Since the data which performed the writing within the slot concerned are not reflected in the magnetic disk drive 5 when it writes in, it does not cancel from the un-volatilizing-ized device 37. And the information stored in the slot management information 1 storing section 32b4 about DESUTEJI within the slot management block 32b3 which manages the slot concerned, the slot management information 2 storing section 32b5, and the slot management information 3 storing section 32b6 is reset (step 207).

[0056] As mentioned above, although this invention was concretely explained based on the example, this invention is not limited to said example and can be variously changed in the range which does not deviate from the summary. For example, it is good also as a configuration equipped with two or more microprocessors 31 in the light after mold disk unit 3, data transfer control circuits 34 for a channel, data transfer control circuits 35 for a disk, etc. Moreover, the configuration of central processing units 1a and 1b, a light after mold disk cache equipment 3, a magnetic disk drive 5, etc., etc. is not limited to what was illustrated in the above-mentioned example.

[0057]

[Effect of the Invention] It will be as follows if the effectiveness acquired by the typical thing among invention indicated in this invention is explained briefly.

[0058] That is, the transfer between cache memory and a host processor can be performed immediately, without according to the disk cache input/output control system which becomes this invention, waiting for termination of DESUTEJI of the data concerned, if the target data are stored in cache memory to the

read-out demand from a host processor. Moreover, when the target data are stored in cache memory and DESUTEJI of the data concerned is completed to the write request from a host processor, the writing to the data made into the purpose can be performed without waiting for termination of the DESUTEJI processing between cache memory and a magnetic disk, and the effectiveness that the fall of a system throughput can be prevented is acquired. Moreover, the high order data transfer way of a between [a host processor and cache memory] and the low order data transfer way of a between [cache memory and a magnetic disk drive] can be used effectively, and the effectiveness that the fall of the availability of a data transfer way can be prevented is acquired.

[Translation done.]

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平5-265860

(43)公開日 平成5年(1993)10月15日

(51)Int.Cl. ⁵	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 12/08	3 2 0	7232-5B		
3/06	3 0 2 A	7165-5B		

審査請求 未請求 請求項の数3(全 12 頁)

(21)出願番号 特願平4-64187

(22)出願日 平成4年(1992)3月19日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(71)出願人 000233055

日立ソフトウェアエンジニアリング株式会社

神奈川県横浜市中区尾上町6丁目81番地

(72)発明者、田中 真由美

神奈川県小田原市国府津2880番地 株式会社日立製作所小田原工場内

(74)代理人 弁理士 富田 和子

最終頁に続く

(54)【発明の名称】 ディスクキャッシュ入出力制御システム

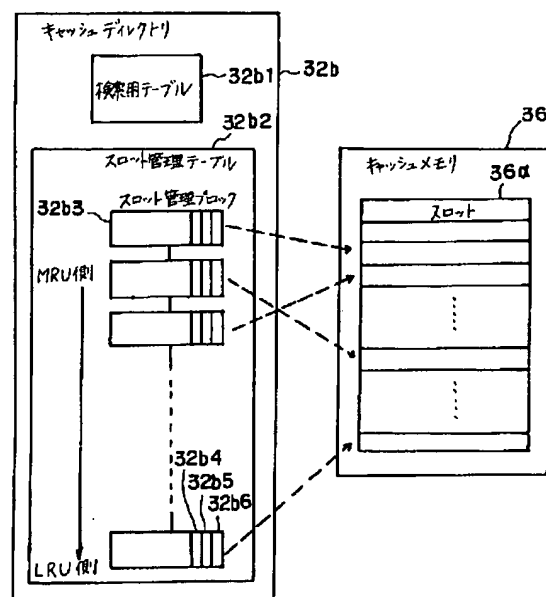
(57)【要約】

【目的】 デステージ中のスロットに対し、上位処理装置から入出力要求が発生した場合、当該スロットのデステージ終了を待たずに、キャッシュメモリと上位処理装置間でのデータ転送を可能とする。

【構成】 キャッシュディレクトリ32b2内のスロット管理テーブル32b3にデステージに関するスロット管理情報1格納部〜3格納部32b4〜32b6を設定することにより、上位チャネル2aまたは2bからアクセス要求のあったデータを含むキャッシュメモリ36内のスロット36aがデステージ中であっても、上位チャネル2aまたは2bからのアクセス要求をデステージと並列に行うことができるディスクキャッシュ入出力制御システム。

【効果】 デステージ終了を待たずに、上位処理装置とキャッシュメモリ間でデータ転送が行えるため、システムのスループットの低下を防止することができる。

図3



【特許請求の範囲】

【請求項1】 回転型記憶装置に格納されているデータの写しを保持する複数のスロットを持つキャッシュメモリを有し、上位処理装置から与えられる前記回転型記憶装置に対する入出力命令の実行に際し、前記キャッシュメモリに保持されているデータを用いて応答するディスクキャッシュ装置において、前記キャッシュメモリと前記上位処理装置との間のデータ転送を行なう第1の転送制御手段と、前記キャッシュメモリと前記回転型記憶装置との間のデータ転送を行なう第2の転送制御手段と、前記キャッシュメモリの各スロットに対応し、対応する各スロットに関する管理情報を格納する記憶手段と、前記記憶手段に前記管理情報を書き込み、前記入出力命令の実行時に前記記憶手段に格納された前記管理情報を参照して、対応するスロットに格納されたデータが前記第2の転送制御手段による前記回転型記憶装置に対する書き込みをさまたげないデータであるかを判定し、書き込みをさまたげないデータについては、前記第1の転送制御手段に転送制御のための指令を発行するプロセッサとを有することを特徴とするディスクキャッシュ入出力制御システム。

【請求項2】 上記請求項1記載のディスクキャッシュ入出力制御システムにおいて、前記キャッシュメモリの各スロットに対応する前記管理情報は、対応する各スロットに格納されたデータを前記回転型記憶装置へ書き込み（デステージ処理）中か否か、および対応スロット中のいずれのレコードに格納されたデータをデステージ中か、および対応スロットに格納されたデータをデステージ中に前記入出力命令の実行により前記データに書き込みが行なわれたか否かの情報であることを特徴とするディスクキャッシュ入出力制御システム。

【請求項3】 上記請求項2記載のディスクキャッシュ入出力制御システムにおいて、前記上位処理装置から与えられる書き込み命令に対するデータが前記キャッシュメモリ内に存在する場合には、前記プロセッサにより前記記憶手段に格納された前記管理情報を読み出して、この管理情報により前記データを前記キャッシュメモリから前記回転型記憶装置へデステージ中であるかを判定し、前記第2の転送制御手段が前記キャッシュメモリから前記回転型記憶装置へのデステージ処理中であっても、デステージ処理が終了されたデータについては直ちに前記キャッシュメモリ内の前記データに対して書き込みを行って、デステージ処理中に前記キャッシュメモリ内の前記データに対して書き込みを行ったという情報を前記管理情報に格納した後、前記上位処理装置に書き込み終了を報告することを特徴とするディスクキャッシュ入出力制御システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は、情報処理装置内のディ

スクサブシステムにおけるディスクキャッシュ入出力制御システムに適用して有効な技術に関する。

【0002】

【従来の技術】 情報処理システムにおいては、中央処理装置は半導体メモリから成る高速な主記憶をもち、近年、その処理速度は著しく向上している。一方、外部記憶装置としては、回転型記憶装置の一種である磁気ディスク装置を使用することが知られている。

【0003】ところで、磁気ディスク装置は、複数枚の記録円盤（これを磁気ディスクと呼ぶ）を重ねた構造になっている。各磁気ディスク上には、データを記録する領域として、同心円上に複数のトラックが設けられており、1つのトラックには複数のレコードが格納される。磁気ディスクの回転軸上から等距離にあるトラックが円盤の枚数だけ垂直に並んでいる。この集合はシリンダと呼ばれている。

【0004】データをアクセスする時には、シリンダ番号、トラック番号、レコード番号を指定すると、各々の磁気ディスクの面上に付けられた磁気ヘッドを所定の駆動装置が、磁気ディスクの径方向に移動させ、任意のトラック上に位置付け、目的のレコードが記録されている、または、レコードが記録されるべき、当該トラック上の領域が磁気ディスクの回転によって、磁気ヘッドの直下を通過する時に、目的のレコードを読み出し、または、レコードの書き込みを行う。

【0005】したがって、磁気ヘッドを目的のトラックに移動させるシーク動作や当該トラック上の目的のレコードが磁気ヘッドの直下を通過するまでの回転待ちなどの機械的動作が介在することとなる。このような機械的動作に要する時間は、中央処理装置の処理時間と比較すると大きいために、磁気ディスクからのデータの入出力時間が、情報処理システムの性能を抑える一つの要因となっている。

【0006】上述のような不都合を回避すべく、上位処理装置と磁気ディスク装置との間のデータ転送経路に、半導体メモリから成るキャッシュメモリを配置し、上位処理装置側から入出力を要求される可能性の高いデータを磁気ディスク装置からキャッシュメモリに予め複写しておき、上位処理装置側からの入出力要求に対し、可能な限り、当該キャッシュメモリに複写されているデータを用いて応答するようにした、所謂、ディスクキャッシュ技術が用いられている。これにより、目的のデータはキャッシュメモリと上位処理装置との間で転送され、磁気ディスク装置に対する機械的動作がなくなり、高速な応答が可能となる。

【0007】一般に、ディスクキャッシュ技術では、上位処理装置からの入力命令に対しては、目的のレコードがキャッシュメモリ内に存在すれば、キャッシュメモリから上位処理装置へデータ転送を行い、応答の高速化を図る。

【0008】一方、上位処理装置からの出力命令に対しては、ライトスルー型ディスクキャッシュとライトアフタ型ディスクキャッシュと呼ばれる二通りの処理方法がある。

【0009】ライトスルー型ディスクキャッシュでは、上位処理装置から転送される出力データは、キャッシュメモリが揮発性であるため、キャッシュメモリと磁気ディスク装置との両方に書き込みを行い、電源遮断時などのデータ喪失を防止する。このため、磁気ディスク装置に対する機械的動作が介在し、応答の高速化は図れない。

【0010】これに対し、ライトアフタ型ディスクキャッシュでは、たとえば、特開昭59-220856等に開示されているように、キャッシュメモリの他に不揮発化機構を具備し、上位処理装置から転送される書き込みデータは、キャッシュメモリと不揮発化機構の両方に書き込んだ時点で書き込み終了を上位処理装置に報告する。書き込みデータは不揮発化機構内にバックアップされているため、電源遮断時などのデータ保証が行え、かつ、書き込み時には、磁気ディスク装置に対する機械的動作が介在せず、応答の高速化が図れる。

【0011】さて、ディスクキャッシュ技術では、目的のデータがキャッシュメモリ内に存在することをヒット、目的のデータがキャッシュメモリ内に存在する確率をヒット率と呼ぶが、キャッシュメモリと上位処理装置間でのデータ転送の割合を可能な限り多くするためには、ヒット率を大きくすることが必要である。

【0012】キャッシュメモリに格納すべきデータを選択するアルゴリズムとしてLRU (Least Recently Used) アルゴリズムが広く知られている。LRUアルゴリズムは、最も最近にアクセスされたデータが今後もアクセスされる可能性が高いと考え、より長く、キャッシュメモリ内に留めようとするアルゴリズムである。

【0013】したがって、上位処理装置から要求されたデータがキャッシュメモリ内に存在しない（これをミスと呼ぶ）場合には、磁気ディスク装置からキャッシュメモリへ、そのデータを含むトラックの一部あるいはトラック全体をコピーする（これをステージングと呼ぶ）。ステージングしようとした時に、キャッシュメモリ内に空き領域がない場合には、LRUアルゴリズムにより、最も古い過去にアクセスされたデータが、今後はアクセスされる可能性が最も低いと考え、キャッシュメモリ内から破棄し、この領域に新たにステージングを行う。

【0014】ところで、万一の停電などの障害時に書き込みデータの喪失を防止するため、キャッシュメモリと同時に不揮発化機構にもデータの写しを格納するライトアフタ型ディスクキャッシュでは、不揮発化機構が小容量であるため、書き込みデータで溢れることがないように、データはキャッシュメモリから磁気ディスク装置に随時、コピーする（これをデステージと呼ぶ）。デステー

ジされたデータは不揮発化機構から破棄される。磁気ディスク制御装置は、空き時間を利用し、前述のLRUアルゴリズムにより、今後アクセスされる可能性の最も低いデータを含むトラック全体もしくは、当該トラックの近傍の複数のトラックを随時、デステージする。または、上位処理装置からの書き込みの要求があり、書き込みデータをキャッシュメモリと不揮発化機構に書き込もうとした時、上述のデステージでは追いつかず、不揮発化機構に既に、ある一定量以上の書き込みデータが格納されていて、溢れる可能性がある場合には、当該書き込みデータを含むトラックとその近傍にある複数のトラックをデステージする。

【0015】このように、小容量の不揮発化機構が書き込みデータで溢れることがないように、磁気ディスク制御装置により、効率的に管理され、書き込まれたデータは随時、キャッシュメモリから磁気ディスク装置にデステージされる。また、上位処理装置からの直接的な命令により、積極的にデステージを行うことも可能である。

【0016】

【発明が解決しようとする課題】上記のような従来技術では、キャッシュメモリは複数のエントリ（このエントリはスロットと呼ばれる）から成り、各エントリに磁気ディスク装置の1トラック分のデータが格納されるが、上位処理装置から入出力命令が発行された時、キャッシュメモリ内に複写されている当該データを含むトラック（スロット）がデステージ中である場合には、当該スロットは使用中（スロットビジーと呼ぶ）となる。このため、上位処理装置からの入出力命令は、当該スロットのデステージが終了するまで待たされることとなる。目的のデータがキャッシュメモリ内に存在している（ヒット）にもかかわらず、キャッシュメモリと上位処理装置間のデータ転送が、デステージにより待たされることとなり、システムのスループット（単位時間当りのデータ転送量）及びデータ転送経路の可用性が低下するという問題があった。

【0017】そこで、本発明の目的は、スループットの低下及びデータ転送経路の可用性の低下を防止するため、上位処理装置から入出力要求のあったデータを含むキャッシュメモリのスロットがデステージ中であっても、キャッシュメモリと上位処理装置間のデータ転送を可能とする入出力制御技術を提供することにある。

【0018】本発明の前記ならびにその他の目的と新規な特徴は、本明細書の記述および添付図面から明らかになるであろう。

【0019】

【課題を解決するための手段】本願において開示される発明のうち、代表的なものの概要を簡単に説明すれば、下記の通りである。

【0020】すなわち、本発明になるディスクキャッシュ入出力制御システムでは、上位処理装置と磁気ディス

ク装置との間に、両者間で授受されるデータのコピーを保持するキャッシュメモリ、特に、ライトアフタ型ディスクキャッシュ（上位処理装置から発行されるデータ書き込み要求によって、磁気ディスク装置に書込むべき書き込みデータをキャッシュメモリに格納する動作と当該キャッシュメモリから磁気ディスク装置に書込む動作を非同期に行うキャッシュメモリ）を設け、上位処理装置から発行される入出力命令の実行に際しては、可能な限り、キャッシュメモリに保持されているデータを用いて応答するようにした情報処理システムであって、キャッシュメモリ内の各スロットに対し、デステージに関するスロット管理情報、

1. 当該スロットをデステージ中か否か
2. 当該スロットをデステージ中の時、当該スロットのどのレコードをデステージ中か
3. 当該スロットをデステージ中に、当該スロットに書き込みを行ったか否か

を記憶する手段を設け、上位処理装置から入出力要求が発行された時に、目的のデータを含むスロットがデステージ中であっても、当該入出力要求を受け付け、上位処理装置とキャッシュメモリ間のデータ転送とデステージとを並列に行うことを可能とするものである。

【0021】

【作用】上記した本発明のディスクキャッシュ入出力制御システムによれば、たとえば、キャッシュメモリ内のスロットに対し、デステージが発生した場合には、前述のスロット管理情報1として、“当該スロットをデステージ中”という情報を記憶する。そして、当該スロット内に保持されているレコードを順次、デステージする際、前述のスロット管理情報2として、デステージを開始するレコードの番号を記憶する。上位処理装置から入出力要求が発行された時には、当該要求が読出し要求か書き込み要求かを判定する。読出し要求であり、目的のデータがキャッシュメモリ内に存在する場合には、目的のデータを含むスロットをデステージ中であるか否かを、上記スロット管理情報1により調べ、デステージ中であっても、目的のデータをキャッシュメモリから上位処理装置に直ちに、転送する。

【0022】一方、上位処理装置からの入出力要求が、書き込み要求であり、目的のデータがキャッシュメモリ内に存在する場合には、上述の読出し要求の場合と同様に、スロット管理情報1により目的のデータを含むスロットをデステージ中であるか否かを調べ、デステージ中である場合には、上記スロット管理情報2により、現在デステージ中のスロット内のレコード番号を調べる。目的のデータを含むレコードのデステージが終了していれば、直ちに、上位処理装置からのデータを当該スロット内の目的のレコードに書き込み、不揮発化機構にも、当該レコードをコピーして、上位処理装置に書き込み終了を報告する。同時に、前述のスロット管理情報3として、

“当該スロットをデステージ中に当該スロットに書き込みを行った”ことを記憶する。これにより、当該スロットのデステージ完了後に、当該スロットが不揮発化機構から破棄されることを防止する。

【0023】なお、上記の3つのスロット管理情報に記憶された情報はデステージ完了後にリセットするものとする。

【0024】このようにすることにより、上位処理装置から入出力要求が発行された場合、目的のデータを含むキャッシュメモリ内のスロットがデステージ中であっても、上位装置からの読み出し要求時および既にデステージが終了しているレコードに対する書き込み要求時においては、キャッシュメモリと磁気ディスク装置間のデータ転送経路である下位バスを用いて行われるデステージと並行して、キャッシュメモリと上位処理装置間のデータ転送経路である上位バスを用いてのデータ転送を行うことができる。このため、データ転送経路が有効に利用され、上位処理装置とキャッシュメモリ間のデータ転送が、デステージ終了を待つことなく、直ちに実行され、システムスループットが向上する。

【0025】

【実施例】以下、図面を参照しながら、本発明の一実施例であるディスクキャッシュ入出力制御システムの一例について詳細に説明する。

【0026】図1は、本発明のディスクキャッシュ入出力制御が実施される情報処理システムの全体構成の一例を示すブロック図である。

【0027】図1に示されるように、本実施例のディスクキャッシュ入出力制御が実施される情報処理システムは、中央処理装置1a、1b、チャネル2a、2b、ライトアフタ型ディスクキャッシュ装置3、磁気ディスク接続装置4および複数の磁気ディスク装置5、上位バス6、下位バス7から成る。

【0028】チャネル2a、2bは中央処理装置1a、1bとライトアフタ型ディスクキャッシュ装置3との間におけるコマンドやデータの授受を制御する。

【0029】磁気ディスク接続装置4は、ライトアフタ型ディスクキャッシュ装置3の指令により、配下の複数の磁気ディスク装置5を適宜選択して、当該ライトアフタ型ディスク制御装置と接続するなどの動作を行う。

【0030】磁気ディスク装置5は、ライトアフタ型ディスクキャッシュ装置3と磁気ディスク接続装置4を介して、前記チャネル2a、2bとの間で授受されるデータを記録/再生する外部記憶媒体である。

【0031】上位バス6は、チャネル2a、2bとライトアフタ型ディスクキャッシュ装置3を接続しており、このバスを介してチャネル2a、2bとライトアフタ型ディスクキャッシュ装置3の間のデータ転送が行なわれる。

【0032】下位バス7は、ライトアフタ型ディスクキ

キャッシュ装置3と磁気ディスク接続装置4を接続しており、このパスを介してライトアプタ型ディスクキャッシュ装置3と磁気ディスク接続装置4の間のデータ転送が行なわれる。

【0033】図2は、図1のライトアプタ型ディスクキャッシュ装置3の詳細な構成の一例を示すブロック図である。

【0034】ライトアプタ型ディスクキャッシュ装置3は、図2に示されるように、マイクロプロセッサ31、制御メモリ32、データ転送制御回路33、対チャネル転送制御回路34、対ディスク転送制御回路35、キャッシュメモリ36および不揮発化機構37から成る。

【0035】マイクロプロセッサ31は、ライトアプタ型ディスクキャッシュ制御装置3全体の制御を行う。制御メモリ32はマイクロプログラム32a、キャッシュディレクトリ32bを格納している。

【0036】データ転送制御回路33は、対チャネル転送制御回路34と対ディスク転送制御回路35との同期をとりつつ、チャネル2a、2bと磁気ディスク装置5との間のデータ授受を制御する。

【0037】対チャネル転送制御回路34は、チャネル2a、2bとの間におけるコマンドやデータなどの授受を制御し、対ディスク転送制御回路35は、磁気ディスク装置5との間におけるコマンドやデータなどの授受を制御する。

【0038】キャッシュメモリ部36は、通常、揮発性の半導体メモリで構成され、複数のスロット36aを有する。各スロット36aには、磁気ディスク装置5に記録されている1トラック分のデータの写しが格納される。

【0039】これに対し、不揮発化機構37は、データを磁気ディスク装置5に確実に書き込むことを保証するため、たとえば、不揮発性の半導体メモリなどで構成され、万一の停電などの障害時にも書き込みデータを確実に保持して、当該書き込みデータの信頼性を確保するためのものである。

【0040】図2の制御メモリ32内のディレクトリメモリ32bは、図3に示すように、検索用テーブル32b1およびスロット管理テーブル32b2から成る。

【0041】検索用テーブル32b1は、キャッシュメモリ36に格納（ステージング）されているデータの、磁気ディスク5内のトラックのアドレスと、そのデータのキャッシュメモリ36内の格納位置を示す情報とを保持している。マイクロプロセッサ31は、この検索用テーブル32b1を参照することにより、中央処理装置1a又は1bから要求されたデータがキャッシュメモリ36内に存在するか否かを判定する。

【0042】スロット管理テーブル32b2は、キャッシュメモリ36の各スロット36aと1対1に対応するエントリを保持する。各エントリは前述のLRUアルゴ

リズムに従って順序付けされ、最も最近アクセスされたデータの位置する側をMRU (Most Recently Used) 側と呼び、最も古い過去にアクセスすることを保証するため、たとえば、不揮発性の半導体メモリなどで構成され、万一の停電などの障害時にも書き込みデータを確実に保持して、当該書き込みデータの信頼性を確保するためのものである。

【0043】制御メモリ32内のディレクトリメモリ32bは、図3に示すように、検索用テーブル32b1およびスロット管理テーブル32b2からなる。検索用テーブル32b1は、キャッシュメモリ36に格納（ステージング）されているデータの磁気ディスク装置5内のトラックのアドレスと、そのデータのキャッシュメモリ36内の格納位置を示す情報を保持している。マイクロプロセッサ31は、この検索用テーブル32b1を参照することにより、中央処理装置1a又は1bから要求されたデータがキャッシュメモリ36内に存在するか否かを判定する。

【0044】スロット管理テーブル32b2は、キャッシュメモリ36の各スロットと1対1に対応するエントリを保持する。各エントリは前述のLRUアルゴリズムに従って順序付けされ、最も最近アクセスされたデータの位置する側をMRU (Most Recently Used) 側と呼び、最も古い過去にアクセスされたデータの位置する側をLRU (Least Recently Used) 側と呼ぶ。

【0045】キャッシュメモリ36内の1つのスロット36aに格納されているデータ（磁気ディスク装置5に記録されている1トラック分のデータ）を破棄する時は、LRU側にあるエントリに対応するトラックが選択される。

【0046】1つのエントリに記憶する情報は、たとえば、対応するスロット36aに格納されているデータの磁気ディスク装置番号、シリンダ番号、トラック番号等である。また、スロット管理ブロック32b3には本発明による前述の3つのスロット管理情報1～3を、スロット管理情報1格納部32b4～スロット管理情報3格納部32b6にそれぞれ格納する。

【0047】キャッシュメモリ36および不揮発化機構37と、磁気ディスク装置5との間のデータの読出し／書き込み動作は、たとえば、磁気ディスク装置5における1トラック分のデータ量などを単位として行われ、通常、上位のチャネル2aまたは2bとの間におけるデータの授受とは非同期に行われる。

【0048】次に、本実施例のディスクキャッシュ入出力制御システム的作用について、図4および図5のフローチャートを参照しながら説明する。図4は、図2のライトアプタ型ディスクキャッシュ装置3のMPU31の指令で、対チャネル転送制御回路34により、上位バス6で行われるデータ転送処理を示すフローチャートであ

る。磁気ディスク装置5に格納されているデータのうち、チャンネル2 a、2 bによるアクセスが予想されるデータ領域がMPU 31によりトラック単位などで、キャッシュメモリ36に随時複写されている。チャンネル2 aまたは2 bからの読出し要求に対しては、可能な限り、当該キャッシュメモリ36に複写されているデータを用いて高速に応答する。チャンネル2 aまたは2 bからの書き込み要求に対しては、キャッシュメモリ36に複写されているデータに書き込みを行い、同時に、不揮発化機構37にも書き込みを行う。この後、直ちに、データ書き込み要求発行元のチャンネル2 aまたは2 bに書き込み終了が報告される。不揮発化機構37に書込まれた複数のデータは、随時、まとめて目的の磁気ディスク装置5にデステージされる。磁気ディスク装置5に書込まれたデータは、不揮発化機構37から破棄され、不揮発化機構37内の当該データ領域が開放され、他の書き込みデータのために使用される。

【0049】ところで、従来は、不揮発化機構37から磁気ディスク装置5にデステージされている時に、上位チャンネル2 aまたは2 bから、現在、デステージ中のデータ領域（スロット）に対し、アクセス要求が発生した場合には、当該スロットが使用中であるため、上位チャンネル2 aまたは2 bからの当該スロットに対するアクセス要求は、デステージが終了するまでは受け付けられない。本実施例では、このような場合、スロット管理テーブル32 b 2内のスロット管理ブロック32 b 3内のスロット管理情報1格納部32 b 4、スロット管理情報2格納部32 b 5、およびスロット管理情報3格納部32 b 6におのおの格納されたスロット管理情報1、2および3を利用して、次のように、デステージ処理と上位チャンネル2 aまたは2 bからのアクセス要求を並列に処理できるように制御する。

【0050】すなわち、図2のライトアプタ型ディスクキャッシュ装置3のMPU 31は、キャッシュディレクタ32内の検索用テーブル32 b 1（図3参照）を調べることにより、上位チャンネル2 aまたは2 bからアクセス要求されたデータの写しが、キャッシュメモリ36内に格納されているか否かを調べる（ステップ101）。目的のデータの写しがキャッシュメモリ36内に格納されていない場合には、従来技術による処理（キャッシュミスの処理）を行う（ステップ110）。目的のデータの写しがキャッシュメモリ36内に格納されている場合には、MPU 31はキャッシュメモリ36内のスロット管理テーブル32 b 2から、当該データの写しを含むスロットを管理しているスロット管理ブロック32 b 3を参照する。MPU 31は、まず、スロット管理情報1格納部32 b 4に格納された情報により、当該スロットがデステージ中であるか否かを調べる（ステップ102）。当該スロットがデステージ中でない場合には、従来技術による処理（キャッシュヒット）の処理を行う

（ステップ107、108、109）。当該スロットがデステージ中である場合には、MPU 31は、上位チャンネル2 aまたは2 bからのアクセス要求が読出しであるか、書込みであるかを判定する（ステップ103）。上位チャンネル2 aまたは2 bからのアクセス要求が読出し要求である場合には、MPU 31の指令により、対チャンネルデータ転送制御回路34が、上位バス6を用いて、直ちに、キャッシュメモリ36内のデステージ中のスロットから読みだし要求発行元の上位チャンネル2 aまたは2 bへデータを転送する（ステップ108）。これにより、デステージ中のスロットに対する上位チャンネル2 aまたは2 bからの読出し要求に対しては、デステージの終了を待つことなく、リードヒットの処理が行なわれる。

【0051】上位チャンネル2 aまたは2 bからのアクセス要求が書き込み要求である場合には、MPU 31は、目的のデータの写しを含むスロット36 aを管理するスロット管理ブロック32 b 3内のスロット管理情報2格納部32 b 5に格納された、当該スロット36 a内のデステージ中のレコード番号（レコード番号aとする）を取出し、上位チャンネル2 aまたは2 bからアクセス要求のあったデータを含む当該スロット36 a内のレコード番号（レコード番号bとする）と比較する（ステップ104）。一般に、デステージはスロット内の先頭レコードから順番に行われるので、レコード番号a > レコード番号bであれば、上位チャンネル2 aまたは2 bからアクセス要求のあったデータは既に、デステージが終了していることとなる。このときは、MPU 31の指令により対チャンネルデータ転送制御回路34による上位バス6を用いて、キャッシュメモリ36内の当該スロットのレコード番号bのデータに直ちに書き込みを行い、同時に、不揮発化機構37にも、同じデータを書込むライトヒット処理を行ない（ステップ105）、データ書き込み要求発行元の上位チャンネル2 aまたは2 bに書き込み終了を報告する。さらに、MPU 31は、デステージ中に、当該スロット36 a内のレコード番号bに対し、書き込みを行ったため、当該スロット36 aのデステージ終了後に、当該スロット内のデータの写しが不揮発化機構37より破棄されることを防止するため、スロット管理ブロック32 b 3内のスロット管理情報3格納部32 b 6に、当該スロットはデステージ中に、新たな書き込みがあったことを記録しておく（ステップ106）。

【0052】このように、デステージ中のスロットに対する上位チャンネル2 aまたは2 bからの書き込み要求は、既に、デステージが終了した部分へのアクセス要求であれば、スロット全体のデステージの終了を待つことなく、直ちに、ライトヒットとして処理できる。ただし、レコード番号a ≤ レコード番号bである場合は、上位チャンネル2 aまたは2 bからアクセス要求のあったデータは、現在、デステージ中かまたは、今後、デステージさ

れるレコードに含まれていることを示すので、従来技術と同様に、当該スロットのデステージ終了を待って書込みを行う（ステップ111）。

【0053】なお、上位チャネル2aまたは2bからのアクセス要求が読出しであるか、書込みであるか、また、上位チャネル2aまたは2bからアクセス要求されたデータがスロット内の何番目のレコードに含まれるかなどは、たとえば、チャネルコマンドであるLocateコマンドなどのパラメータとして与えられることが広く知られている。

【0054】一方、対ディスク転送制御回路35による下位バス7でのデステージ処理について、図5のフローチャートを用いて説明する。図3のMPU31は、現在、デステージが行われているキャッシュメモリ36のスロットについて、当該スロットのデステージの終了および、デステージを行うレコード番号などを、随時、監視する。まず、1スロット分のデータのデステージが終了したか否かを調べる（ステップ201）。1スロット分のデータのデステージが終了していない場合には、スロット内の任意の1レコードのデステージが終了したか否かを調べる（ステップ202）。1レコード分のデステージが終了していない場合には、そのままデステージを続行させる（ステップ204）。1レコード分のデステージが終了した場合には、次にデステージを開始する当該スロット内のレコード番号を、キャッシュメモリ36内のスロット管理テーブル32b2に登録されている当該スロットを管理しているスロット管理ブロック32b3のスロット管理情報2格納部32b5に記録し（ステップ203）、対ディスクデータ転送制御回路35により当該レコードのデステージを行う（ステップ204）。

【0055】1スロット分のデステージが終了した場合には、MPU31は当該スロットを管理しているスロット管理ブロック32b3のスロット管理情報3格納部32b6を参照し、当該スロットをデステージ中に、上位チャネル2aまたは2bからの書込み要求により、当該スロットに対し、書込みを行ったか否かを調べる（ステップ205）。書込みを行わなかった場合には、当該スロットのデータの写しを不揮発化機構37から破棄する（ステップ206）。書込みを行った場合には、当該スロット内の書込みを行ったデータが、磁気ディスク装置5には反映されていないので、不揮発化機構37からは破棄しない。そして、当該スロットを管理するスロット管理ブロック32b3内のデステージに関するスロット管理情報1格納部32b4、スロット管理情報2格納部32b5、スロット管理情報3格納部32b6に格納された情報をリセットする（ステップ207）。

【0056】以上、本発明を実施例に基づき具体的に説明したが、本発明は、前記実施例に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能であ

る。たとえば、ライトアフタ型ディスク装置3におけるマイクロプロセッサ31、対チャネルデータ転送制御回路34および対ディスクデータ転送制御回路35などを複数個、備えた構成としてもよい。また、中央処理装置1a、1bやライトアフタ型ディスクキャッシュ装置3、磁気ディスク装置5などの構成は、前述の実施例に例示したものに限定されない。

【0057】

【発明の効果】本発明において開示される発明のうち、代表的なものによって得られる効果を簡単に説明すれば、以下の通りである。

【0058】すなわち、本発明になるディスクキャッシュ入出力制御システムによれば、上位処理装置からの読み出し要求に対し、目的とするデータがキャッシュメモリ内に格納されていれば、当該データのデステージの終了を待つことなく、直ちにキャッシュメモリと上位処理装置との間での転送が実行できる。また、上位処理装置からの書き込み要求に対し、目的とするデータがキャッシュメモリ内に格納されており、当該データのデステージが終了している場合には、キャッシュメモリと磁気ディスク間のデステージ処理の終了を待たずに、目的とするデータへの書き込みを行なうことができ、システムスループットの低下を防止することができるという効果が得られる。また、上位処理装置とキャッシュメモリ間との上位データ転送路および、キャッシュメモリと磁気ディスク装置間との下位データ転送路を有効に利用でき、データ転送路の可用性の低下を防止することができるという効果が得られる。

【図面の簡単な説明】

【図1】本発明のディスクキャッシュ入出力制御システムの全体構成の一例を示すブロック図。

【図2】図1のライトアフタ型ディスクキャッシュ装置3の詳細な構成の一例を示すブロック図。

【図3】図2のキャッシュディレクトリ32bの詳細な構成と、キャッシュメモリ36のスロット36aとの関係の一例を示すブロック図。

【図4】本発明の一実施例であるディスクキャッシュ入出力制御システムの上位バスによるデータ転送の一例を示すフローチャート。

【図5】本発明の一実施例であるディスクキャッシュ入出力制御システムの下位バスによるデステージ処理の一例を示すフローチャート。

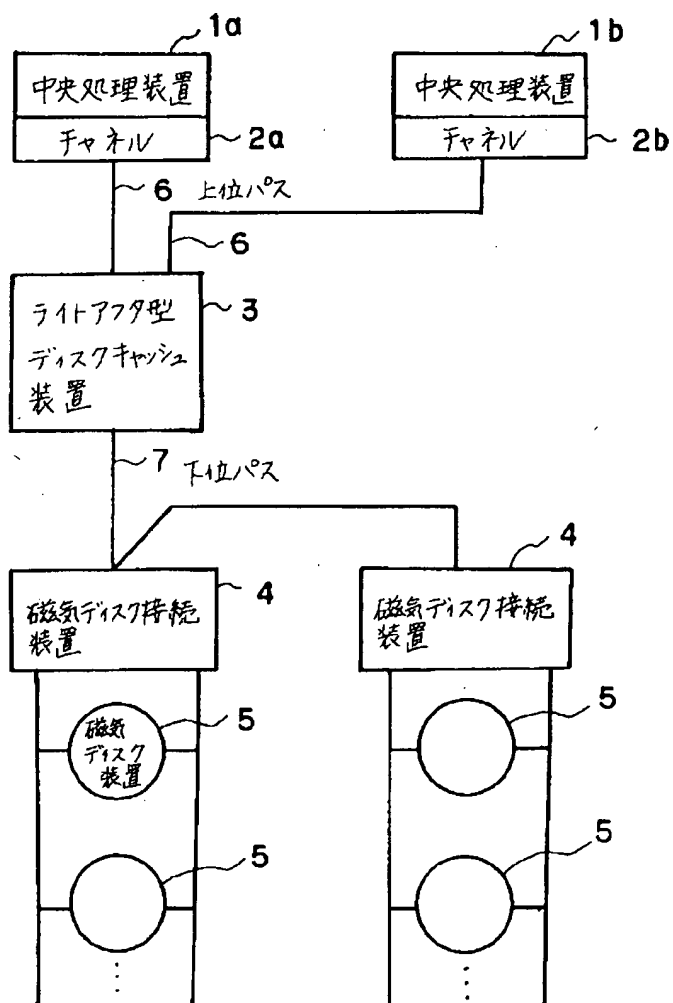
【符号の説明】

- 1a, 1b 中央処理装置
- 2a, 2b チャネル
- 3 ライトアフタ型ディスクキャッシュ制御装置
- 4 磁気ディスク接続装置
- 5 磁気ディスク装置
- 6 上位バス
- 7 下位バス

- | | | | |
|------|--------------|------|----------------|
| 31 | マイクロプロセサ | 32b5 | スロット管理情報2格納部 |
| 32 | 制御メモリ | 32b6 | スロット管理情報3格納部 |
| 32a | マイクロプログラム | 33 | データ転送制御回路 |
| 32b | キャッシュディレクトリ | 34 | 対チャネルデータ転送制御回路 |
| 32b1 | 検索用テーブル | 35 | 対ディスクデータ転送制御回路 |
| 32b2 | スロット管理テーブル | 36 | キャッシュメモリ |
| 32b3 | スロット管理ブロック | 36a | スロット |
| 32b4 | スロット管理情報1格納部 | 37 | 不揮発化機構 |

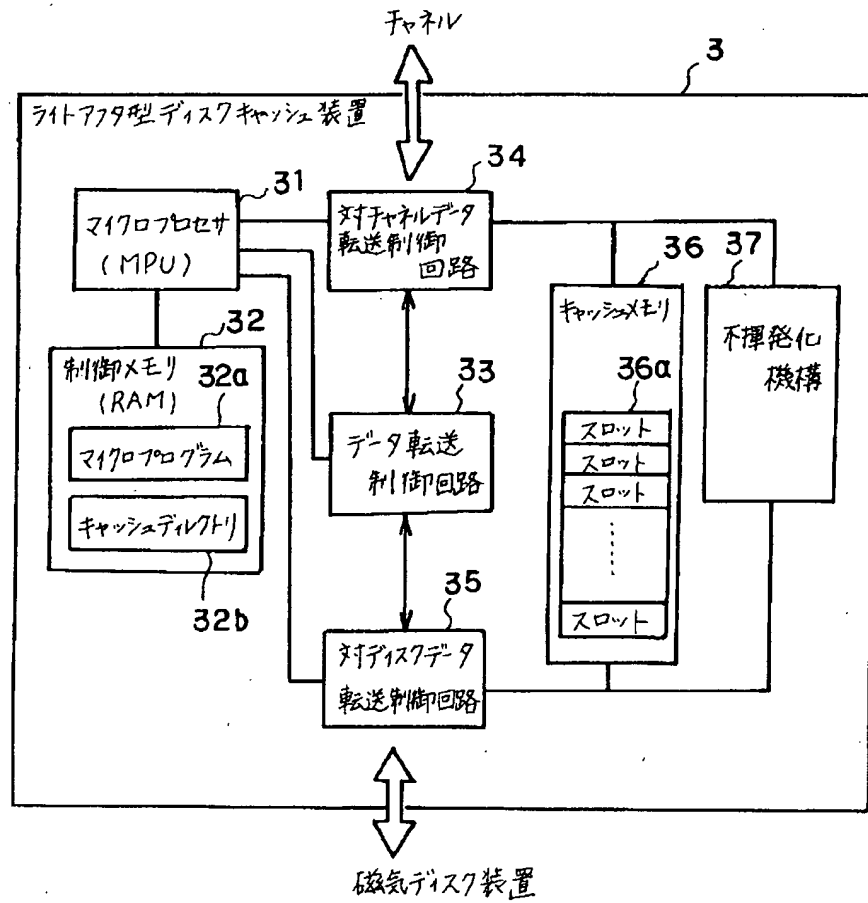
【図1】

図1



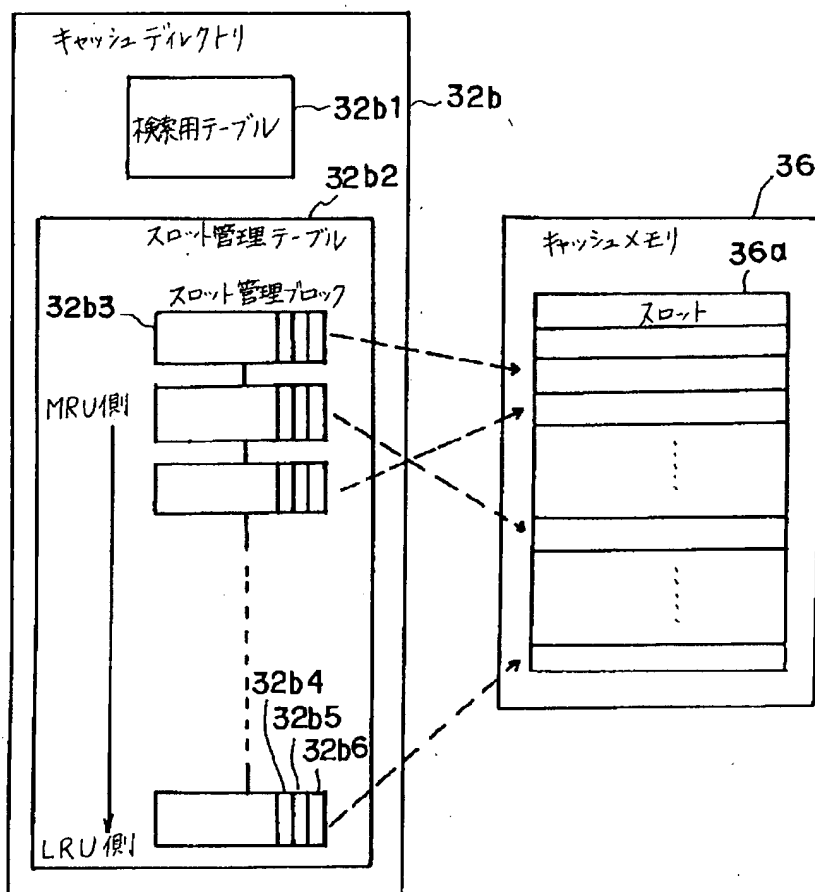
【図2】

図2



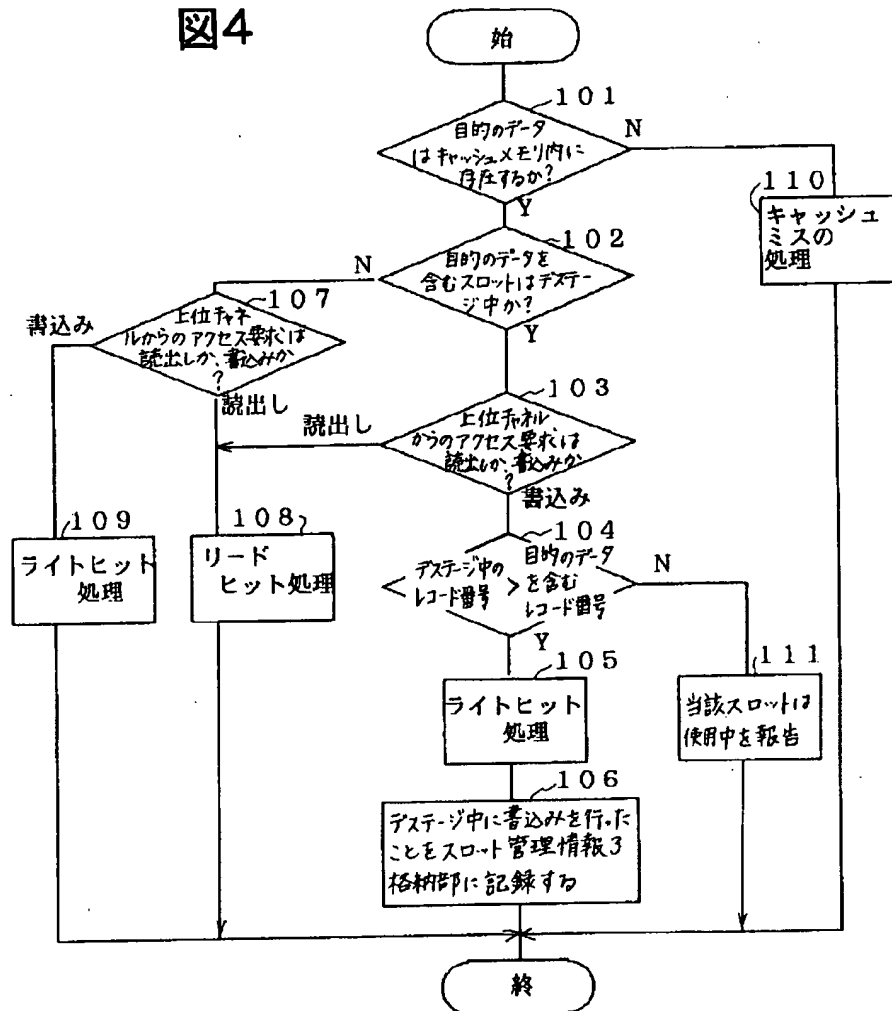
【図3】

図3



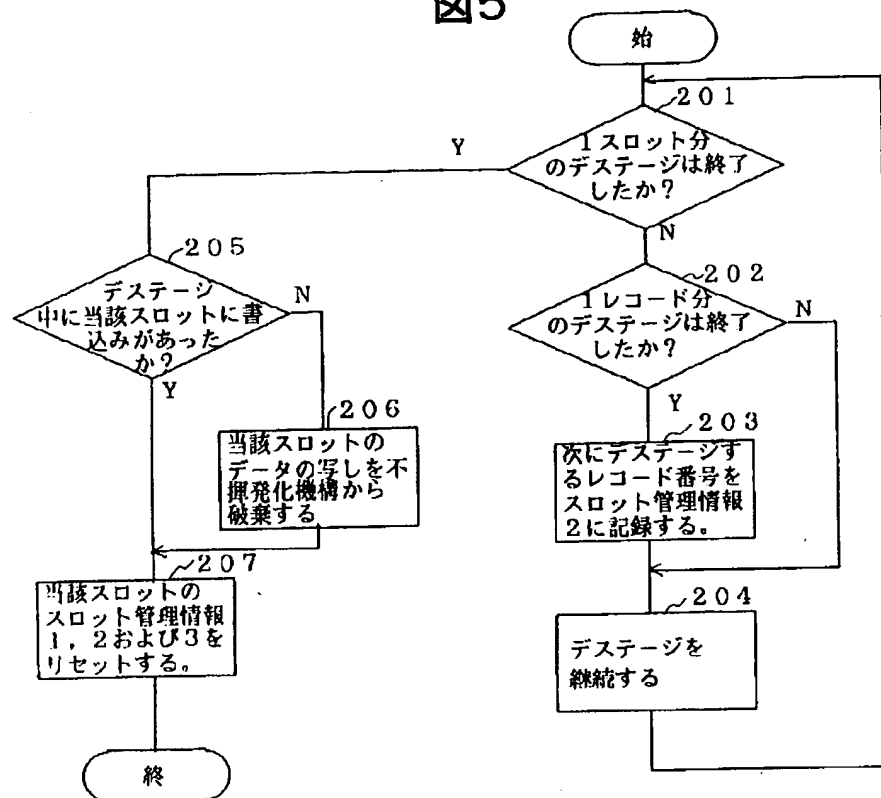
【図4】

図4



【図5】

図5



フロントページの続き

(72)発明者 井坂 昇二
神奈川県横浜市中区尾上町六丁目81番地
日立ソフトウェアエンジニアリング株式会
社内

(72)発明者 森 栄樹
神奈川県横浜市中区尾上町六丁目81番地
日立ソフトウェアエンジニアリング株式会
社内

(72)発明者 四谷 守彦
神奈川県小田原市国府津2880番地 株式会
社日立製作所小田原工場内

(72)発明者 本間 繁雄
神奈川県小田原市国府津2880番地 株式会
社日立製作所小田原工場内